Regression Analysis on Stock Price

Statistical Data Analysis Project

Prakruti Makwana BS20DMU001 BDS Sem 2

CONTENTS

1.INTRODUCTION

1.1 Overview of stocks

- Stock prices are an essential part of the economic activity. In a financial system where the stock market is increasing, we can say that the economy of the company is flourishing.
- Often the stock market is measured as the principal pointer of a country's financial power and progress. Stock market research is important if one wants to earn a major return on stocks as successful forecast of a stock's future value will result in profits for the company.

1.2 Objective of the Analysis

- This report determines the independent variables to predict the opening price of the stock market for the Apple company. Apple Inc. is an American multinational technology company that designs, develops, and sells consumer electronics, computer software, and online services. The analysis focuses on the factors which have the most impact on the opening price.

1.3 Problem statement

- This dataset contains 7 variables: date, open, high, low, close, volume and opening price for each year from 2016 till 2021(present). The response variable is the opening price.
- In this statistical report, correlation between the opening price and other independent variables is explored.
- A sample of 200 datapoints have been taken for analysis. Data has been analysed using JMP software and Microsoft Excel.

2. DATASET EXPLORATION

2.1 Data set dictionary

The data on the following variables has been collected daily from Jan 1st 2016 till April 30th 2021.

- **Opening Price**: The opening price is the value that each share has when the stock exchange opens for trading. The opening price gives a good indication of where the stock will move during the day.

- **Close Price**: Close refers to the price of an individual stock when the stock exchange closed shop for the day. It represents the last buy-sell order executed between two traders.

- **High price**: It is the highest price at which a stock is traded during the day.

- **Low price:** It is the lowest price at which the stock is traded during the day.

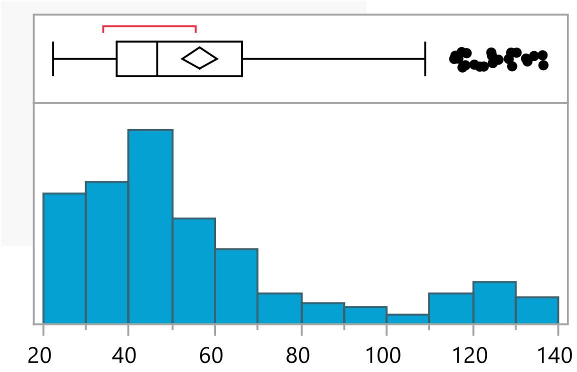
- **Volume:** It is the total number of shares traded in a security over a period. Whenever buyers and sellers exchange shares, the amount gets added to the total volume.

- **Adjusting closed Price:** It is a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions such as dividend payments, stock splits etc that occurred at any time prior to the next days open.

2.2 Descriptive Statistics Measures

Below are the measures for each variable.

1. Open Price (Response variable)



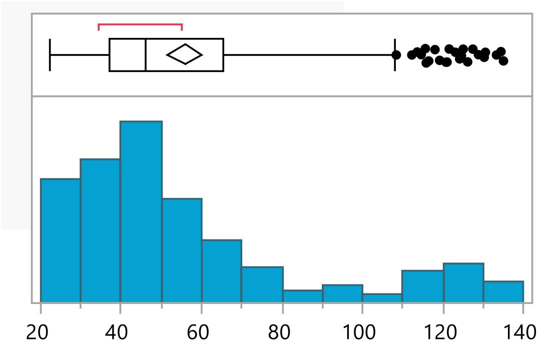
Quantiles

100.0%	maximum	136.479996
75.0%	quartile	66.208126
50.0%	median	46.76375
25.0%	quartile	37.155625
0.0%	minimum	22.5

Summary Statistics

Mean	56.427012
Std Dev	29.925464
Std Err Mean	2.1160498
Upper 95% Mean	60.599771
Lower 95% Mean	52.254254
N	200
Sum	11285.402
Variance	895.53338
Skewness	1.3326713
Kurtosis	0.8168353
Minimum	22.5
Maximum	136.48
Median	46.76375
Mode	39.375

2. Low price



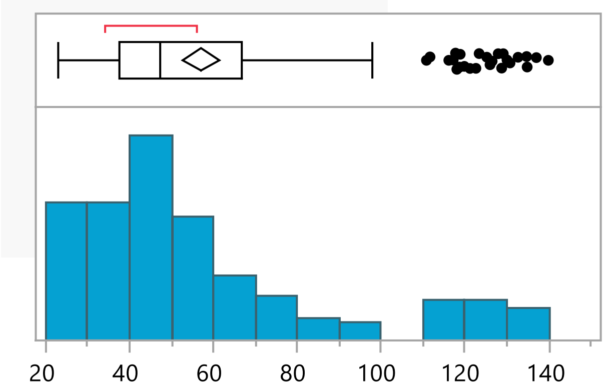
Quantiles

100.0%	maximum	135.020004
75.0%	quartile	65.4937495
50.0%	median	46.3324985
25.0%	quartile	36.97625025
0.0%	minimum	22.3675

Summary Statistics

Mean	55.814175
Std Dev	29.478875
Std Err Mean	2.0844713
Upper 95% Mean	59.924662
Lower 95% Mean	51.703688
N	200
Sum	11162.835
Variance	869.00409
Skewness	1.3281063
Kurtosis	0.8006914
Minimum	22.3675
Maximum	135.02
Median	46.332499
Mode	29.195

3. High price



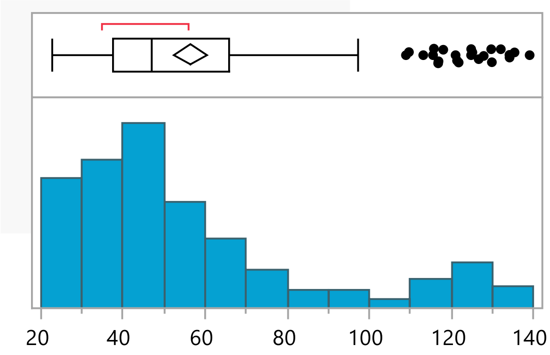
Quantiles

100.0%	maximum	139.850006
75.0%	quartile	66.5875
50.0%	median	47.2625005
25.0%	quartile	37.51625025
0.0%	minimum	22.9175

Summary Statistics

Mean	57.0084
Std Dev	30.226081
Std Err Mean	2.1373067
Upper 95% Mean	61.223076
Lower 95% Mean	52.793724
N	200
Sum	11401.68
Variance	913.61598
Skewness	1.3207304
Kurtosis	0.7842764
Minimum	22.9175
Maximum	139.85001
Median	47.262501
Mode	40

4. Close price



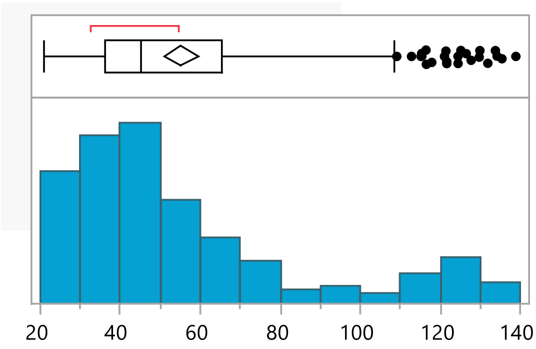
Quantiles

100.0%	maximum	139.070007
75.0%	quartile	66.00375175
50.0%	median	46.7924995
25.0%	quartile	37.4381235
0.0%	minimum	22.584999

Summary Statistics

Mean	56.4248
Std Dev	29.842678
Std Err Mean	2.110196
Upper 95% Mean	60.586015
Lower 95% Mean	52.263585
N	200
Sum	11284.96
Variance	890.58545
Skewness	1.3248941
Kurtosis	0.8038734
Minimum	22.584999
Maximum	139.07001
Median	46.7925
Mode	121.78

5. Adjusting close price



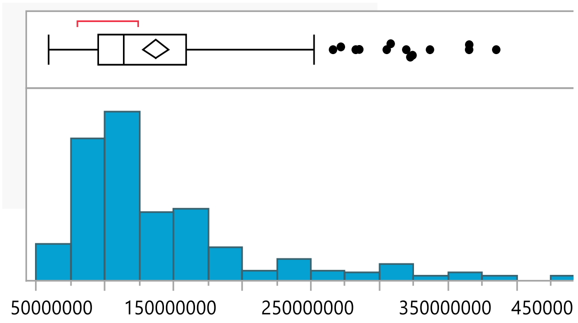
Quantiles

100.0%	maximum	138.862503
75.0%	quartile	65.340977
50.0%	median	45.4342385
25.0%	quartile	36.17768375
0.0%	minimum	21.134403

Summary Statistics

Mean	55.309575
Std Dev	30.298578
Std Err Mean	2.142433
Upper 95% Mean	59.53436
Lower 95% Mean	51.08479
N	200
Sum	11061.915
Variance	918.00385
Skewness	1.3154851
Kurtosis	0.7760771
Minimum	21.134403
Maximum	138.8625
Median	45.434239
Mode	121.59829

6. Volume



Quantiles

100.0%	maximum	447940000
75.0%	quartile	159341700
50.0%	median	114061000
25.0%	quartile	94654600
0.0%	minimum	58676400

Summary Statistics

Mean	137512456
Std Dev	67339177
Std Err Mean	4761598.9
Upper 95% Mean	146902122
Lower 95% Mean	128122790
N	200
Sum	2.75e+10
Variance	4.535e+15
Skewness	1.933331
Kurtosis	4.104803
Minimum	58676400
Maximum	447940000
Median	114061000
Mode	.

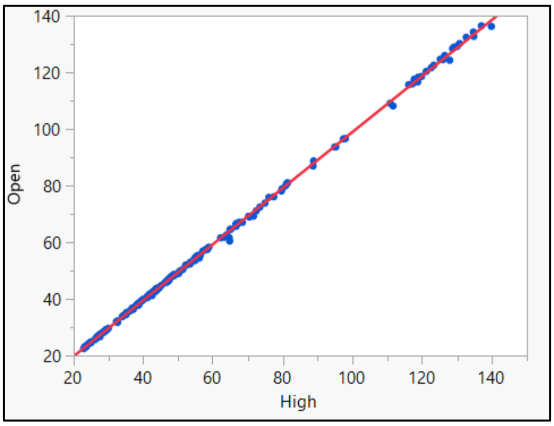
3. Linear Regression Analysis

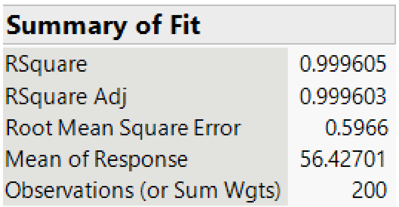
- Linear regression is used to predict the value of a response based on the value of one continuous variable. The method of least squares is used to find the best-fitting line for the observed data which is used to make the prediction of the response variable.
- It performs operations on a dataset where the target values have been defined already. After regression analysis, the variables which don't have an effect on the response variable will not be taken into consideration while fitting the model.

$$Line\ of\ fit:\ \ Y_i = \alpha + \beta x_i + \epsilon_i$$

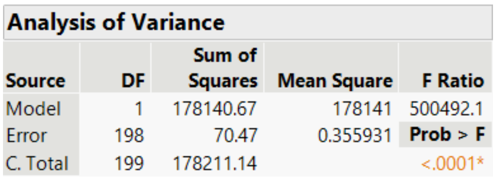
3.1 Line of regression and Plots

Linear regression between Opening price and High Price



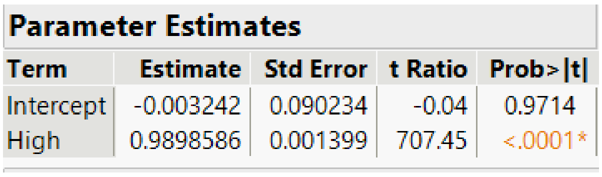


Summary of Fit	
RSquare	0.999605
RSquare Adj	0.999603
Root Mean Square Error	0.5966
Mean of Response	56.42701
Observations (or Sum Wgts)	200



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	178140.67	178141	500492.1
Error	198	70.47	0.355931	Prob > F
C. Total	199	178211.14		<.0001*



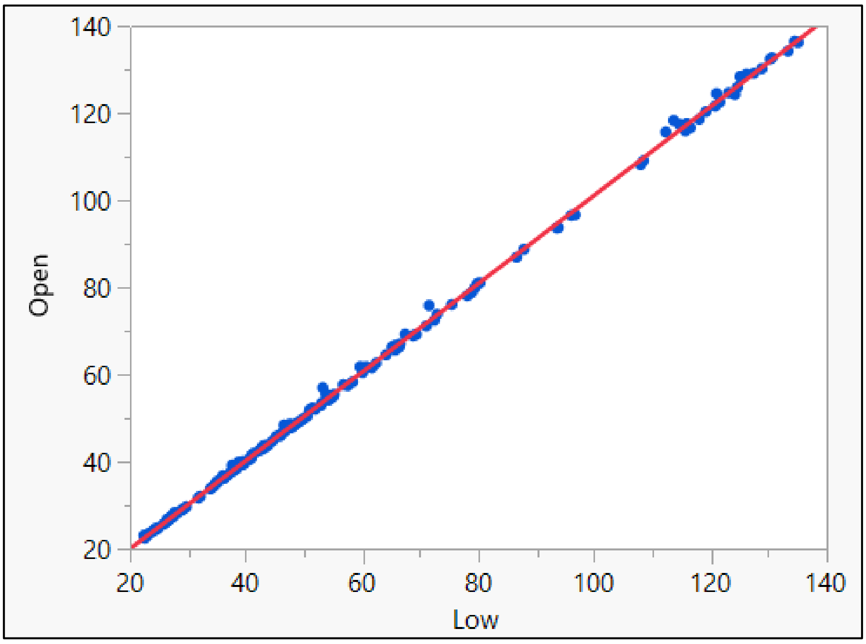
Parameter Estimates

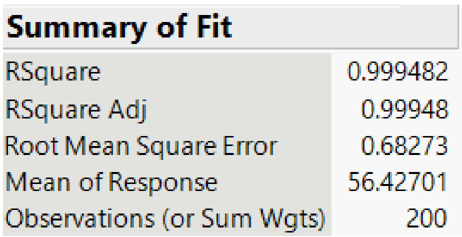
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.003242	0.090234	-0.04	0.9714
High	0.9898586	0.001399	707.45	<.0001*

Observation:

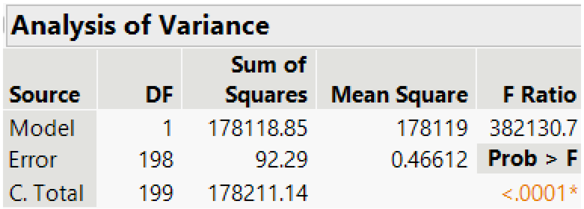
- Line of regression: Opening price = -0.003242 + 0.9898586*High
- The p value (Prob |t|) is less than 0.05, thus we can say that high price is a significant factor for predicting opening price.

Linear regression between Opening price and Low price



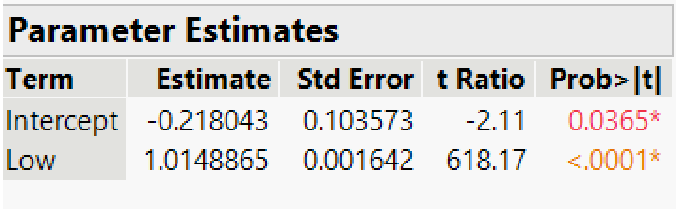


Summary of Fit	
RSquare	0.999482
RSquare Adj	0.99948
Root Mean Square Error	0.68273
Mean of Response	56.42701
Observations (or Sum Wgts)	200



Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	1	178118.85	178119	382130.7
Error	198	92.29	0.46612	Prob > F
C. Total	199	178211.14		<.0001*



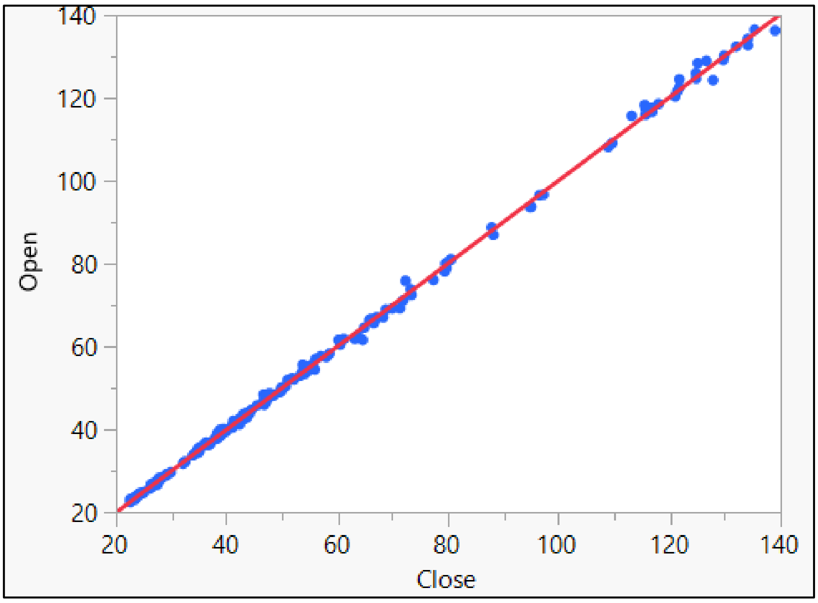
Parameter Estimates

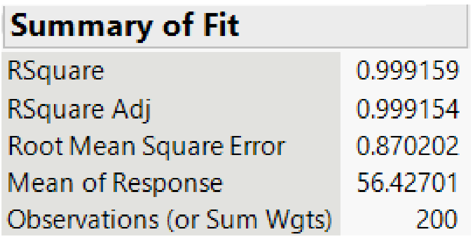
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.218043	0.103573	-2.11	0.0365*
Low	1.0148865	0.001642	618.17	<.0001*

Observation:

- Open = -0.218043 + 1.0148865*Low
- The p value (Prob |t|) is less than 0.05, thus we can say that low price is a significant factor for predicting opening price.

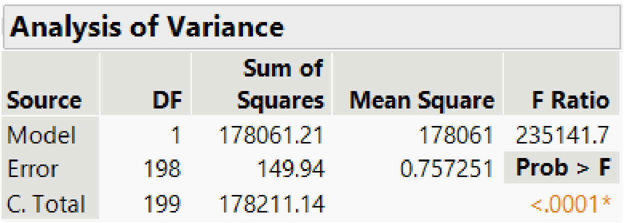
Linear regression between Opening price and Close price





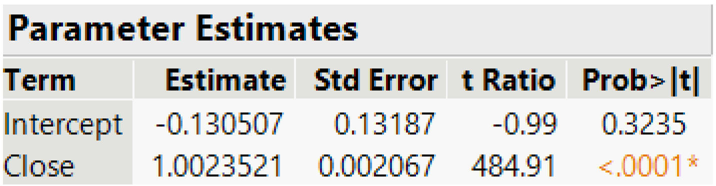
Summary of Fit

RSquare	0.999159
RSquare Adj	0.999154
Root Mean Square Error	0.870202
Mean of Response	56.42701
Observations (or Sum Wgts)	200



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	178061.21	178061	235141.7
Error	198	149.94	0.757251	Prob > F
C. Total	199	178211.14		<.0001*



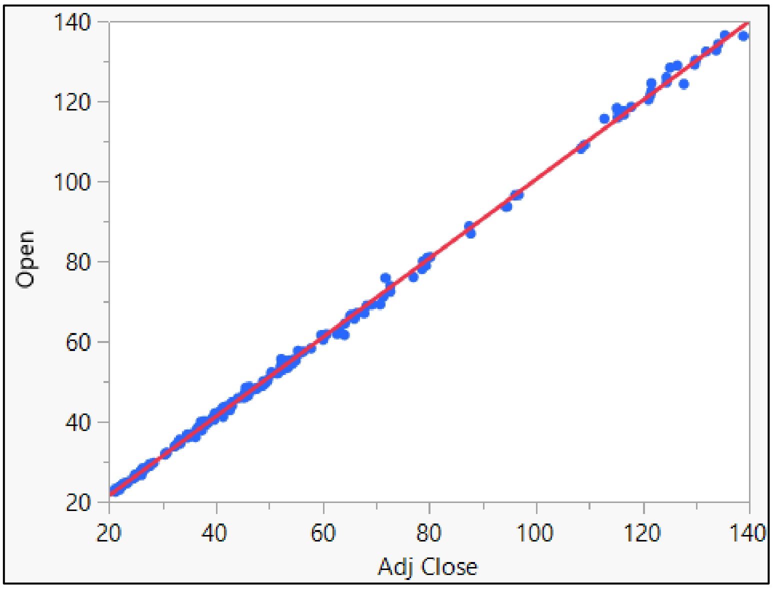
Parameter Estimates

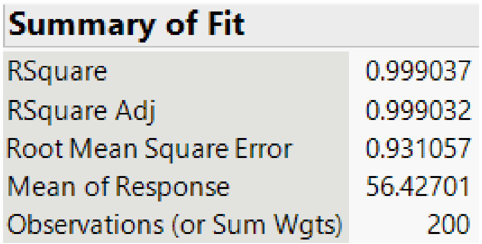
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.130507	0.13187	-0.99	0.3235
Close	1.0023521	0.002067	484.91	<.0001*

Observation:

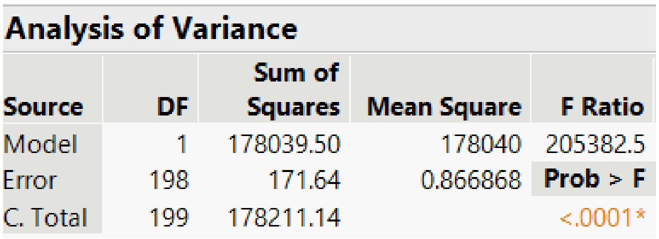
- Open = -0.130507 + 1.0023521*Close
- The p value (Prob |t|) is less than 0.05, thus we can say that close price is a significant factor for predicting opening price.

Linear regression between Open price and Adjusting Close price



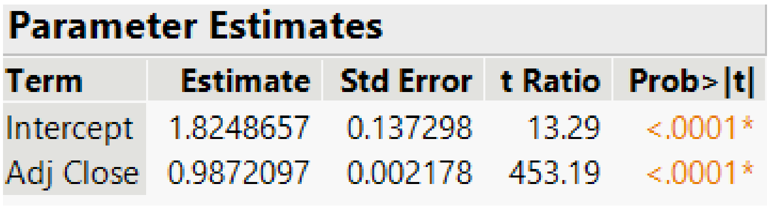


Summary of Fit	
RSquare	0.999037
RSquare Adj	0.999032
Root Mean Square Error	0.931057
Mean of Response	56.42701
Observations (or Sum Wgts)	200



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	178039.50	178040	205382.5
Error	198	171.64	0.866868	Prob > F
C. Total	199	178211.14		<.0001*



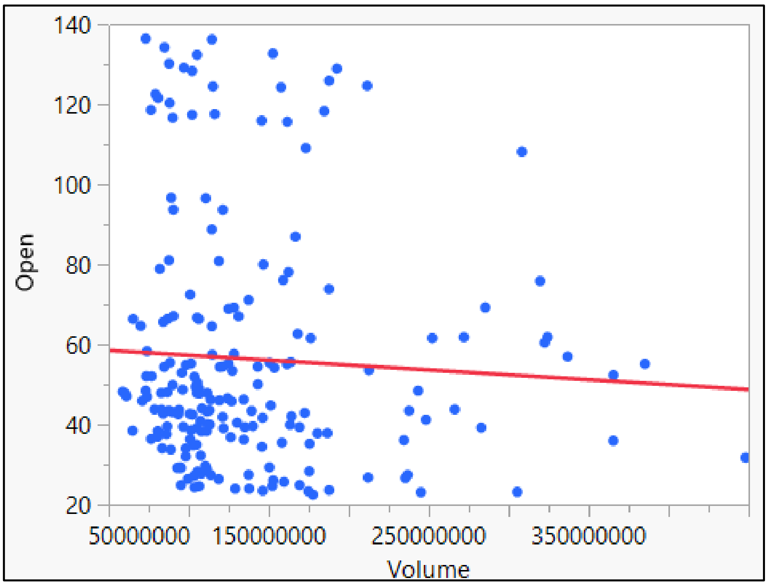
Parameter Estimates

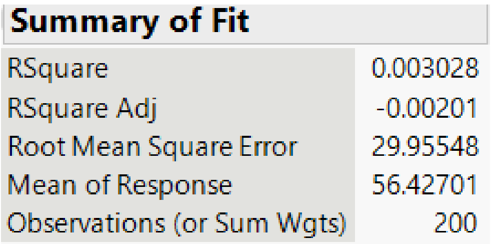
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.8248657	0.137298	13.29	<.0001*
Adj Close	0.9872097	0.002178	453.19	<.0001*

Observation:

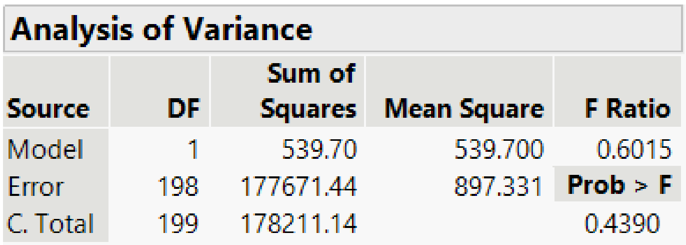
- Open = 1.8248657 + 0.9872097*Adj Close
- The p value (Prob |t|) is less than 0.05, thus we can say that adjusting close price is a significant factor for predicting opening price.

Linear regression between Opening price and Volume



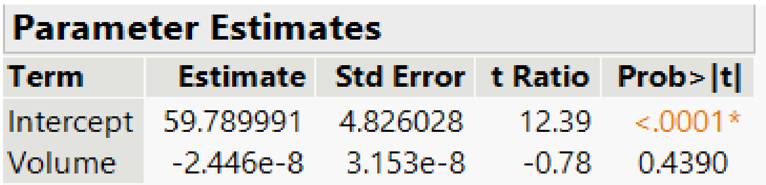


Summary of Fit	
RSquare	0.003028
RSquare Adj	-0.00201
Root Mean Square Error	29.95548
Mean of Response	56.42701
Observations (or Sum Wgts)	200



Analysis of Variance

		Sum of		
Source	**DF**	**Squares**	**Mean Square**	**F Ratio**
Model	1	539.70	539.700	0.6015
Error	198	177671.44	897.331	**Prob > F**
C. Total	199	178211.14		0.4390



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	59.789991	4.826028	12.39	<.0001*
Volume	-2.446e-8	3.153e-8	-0.78	0.4390

Observation:

- Open = 59.789991 - 2.4456e-8*Volume
- The p value (Prob |t|) is more than 0.05, thus we can say that volume is not a significant factor for predicting opening price. Hence, we will drop this variable and not include it for fitting the model.

3.2 Hypothesis Testing

- Assuming that the opening price of years 2020 and 2019 are of unequal variance, then it is to be tested whether sample means can be used as the point estimator for the population mean, at 5% significance level.

T-TEST: TWO-SAMPLE ASSUMING UNEQUAL VARIANCES

	2019	*2020*
MEAN	52.15351239	95.26767
VARIANCE	78.04754423	484.5241
OBSERVATIONS	200	253
HYPOTHESIZED MEAN DIFFERENCE	0	
DF	347	
T STAT	-28.39561024	
P(T<=T) ONE-TAIL	8.07419E-93	
T CRITICAL ONE-TAIL	1.649256711	
P(T<=T) TWO-TAIL	1.61484E-92	
T CRITICAL TWO-TAIL	1.966824003	

$$H_0: Opening\ price(2019) = Opening\ price(2020)$$

$$H_1: Opening\ price(2019) \neq Opening\ price(2020)$$

Range of T critical value: (-1.966, 1.966)

$$T_{Calc} = \frac{\overline{(x_1} - \overline{x_2}) - \mu}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{2}}}$$

$$T_{Calc}: -28.39$$

$$Since, T_{Calc}\ does\ not\ lie\ within\ the\ range, we\ reject\ Null\ hypothesis.$$

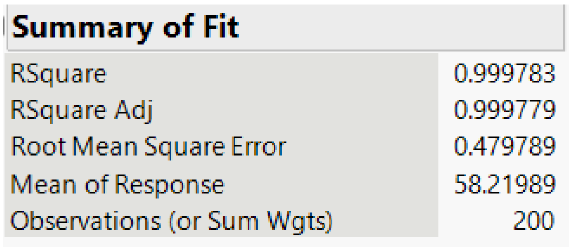
Also, the p value= 1.614E-92 which is less than α=0.05

$$\therefore We\ reject\ H_0$$

4. Multiple regression

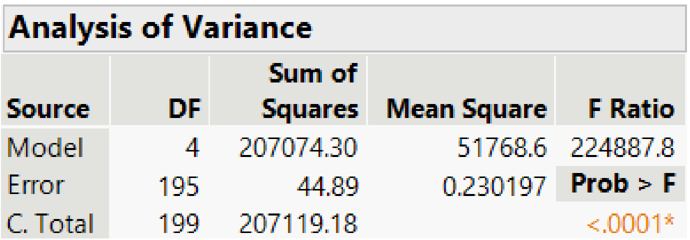
- Multiple linear regression is used to model the relationship between a continuous response variable and continuous or categorical explanatory variables.
- It accommodates more than one predictive factor: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon$ where Y is an observed value of the response variable, each x_i is an observed value for a distinct factor variable, and each of the βs is a parameter of the model and ε is the amount by which an individual response deviates from the model.

4.1 Line of Multiple regression



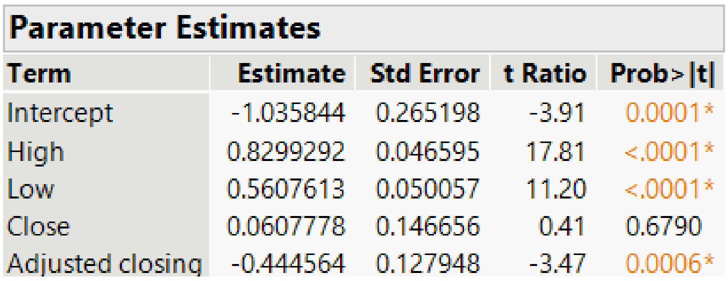
Summary of Fit

RSquare	0.999783
RSquare Adj	0.999779
Root Mean Square Error	0.479789
Mean of Response	58.21989
Observations (or Sum Wgts)	200



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	207074.30	51768.6	224887.8
Error	195	44.89	0.230197	Prob > F
C. Total	199	207119.18		<.0001*



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.035844	0.265198	-3.91	0.0001*
High	0.8299292	0.046595	17.81	<.0001*
Low	0.5607613	0.050057	11.20	<.0001*
Close	0.0607778	0.146656	0.41	0.6790
Adjusted closing	-0.444564	0.127948	-3.47	0.0006*

Observation:

After fitting the model, it is observed that the variable close price is no longer a significant predictor for opening price as the p value is greater than 0.05.

Regression line:

$$Y = -1.035 + 0.829 * high + 0.560 * low + 0.060 * close - 0.444 * adj\ close$$

4.2 Anova

Evaluating the variation between the 5 groups (2016,2017,2018,2019,2020) measures. We want to decide if these 5 groups are different from each other or whether they are the same.

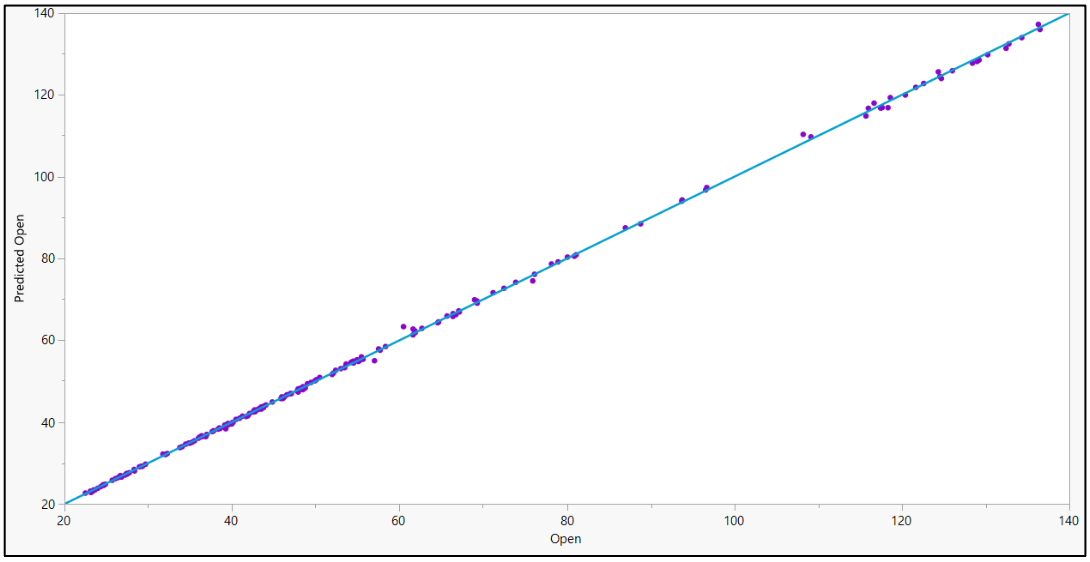
ANOVA: SINGLE FACTOR

SUMMARY						
GROUPS (OPENING PRICE)	*Count*	*Sum*	*Average*	*Variance*		
2016	120	2640.757	26.40757	4.027374		
2017	150	5617.723	37.70284	14.33694		
2018	150	7044.908	47.28126	26.62755		
2019	120	6181.322	52.38409	84.39299		
2020	150	14383.61	95.89072	502.2607		
ANOVA						
SOURCE OF VARIATION	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
BETWEEN GROUPS	385953.5089	4	96488.38	699.5418	7.5E-236	2.385409
WITHIN GROUPS	91172.27347	661	137.9308			
TOTAL	477125.7824	665				

- We see that there are five groups (Opening prices):
- H_0: All groups are equal.
- H_1 : Atleast one is different from the other.
- $F_{Calc} = 699.5418$
- $F_{critical} = 2.3835$
- Therefore $F_{Calc} > F_{critical}$ so we do not accept the null hypothesis at 95% confidence level.
- Also, p-value = 7.5E-236 which is less than the significant level i.e., 0.05, thus we reject Null Hypothesis.

5. Graphs

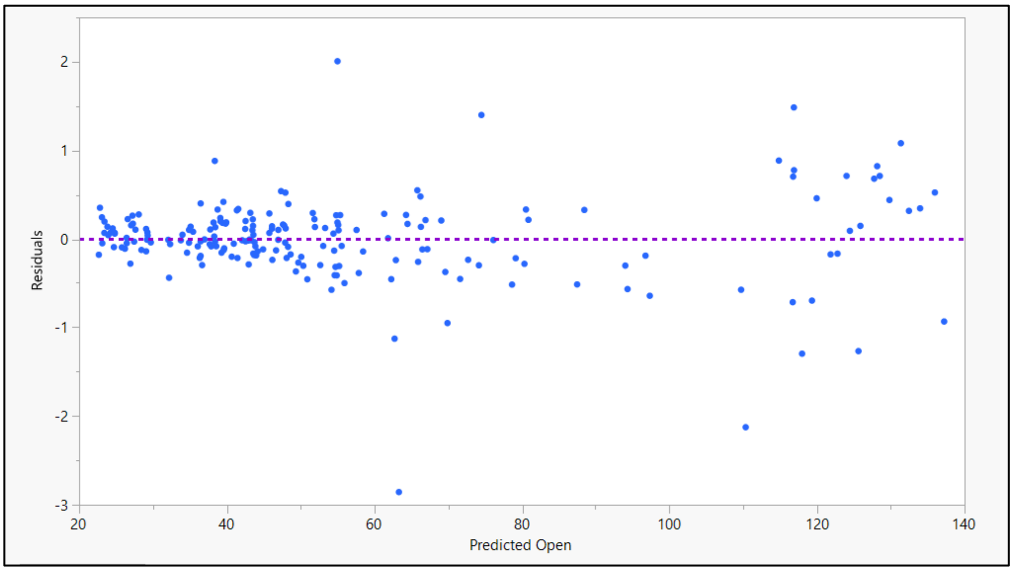
5.1 Scatter plot



Scatter plot of Predicted open price by Open price

- A predicted against actual plot shows the effect of the model and compares it against the null model. This model is accurate with R^2 value equal to 0.999, which is a good fit since the points are close to the fitted line, with narrow confidence bands.

5.2 Residual Plot



Residuals by Predicted open

- A residual is the vertical distance between a data point and the regression line. It gives the difference between the measured value and the predicted value of a regression model

CONCLUSION

- We have seen that regression is a general collection of techniques that are used to model a response as a function of predictor variables. These relations can make a pattern which is used to evaluate trends, make estimates and forecasts.

- With the significant features fitted (Close price, high price, low price, adjusting close), the R^2 value obtained was 0.999. This indicates that these explanatory variables are essential for predicting the response variable, Open price accurately.

- The mean differences between the opening prices of consecutive years shows that the values of stocks has been increasing over the years.

- This report gives only a part of work by focussing on the statistical area, there are critical decisions which have to be made and other key factors which are also responsible to predict the price and forecast the trade.

References

[1]Yahoo Finance NASDAQ. (2021, April 31). Apple Inc (APPL) https://in.finance.yahoo.com/quote/AAPL/history?p=AAPL

[2]: Carver, Robert. 2019. Practical Data Analysis with JMP®, Third Edition. Cary, NC: SAS Institute Inc.

[3] Seethalakshmi, Ramaswamy. (2018). Analysis of stock market predictor variables using linear regression. International Journal of Pure and Applied Mathematics. 119. 369-377.